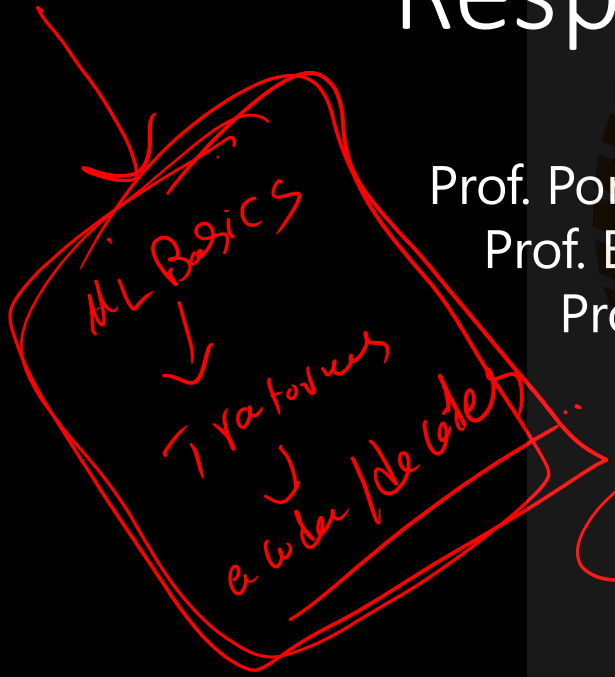
Responsible & Safe AI

Prof. Ponnurangam Kumaraguru (PK), IIITH
Prof. Balaraman Ravindran, IIT Madras
Prof. Arun Rajkumar, IIT Madras



{Machine & Graph} Unlearning

Prompt: A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!

What is happening here?

Prompt: A map of the United States made out of sushi. It is on a table next to a glass of red wine.

Prompt: A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!

What is happening here?

Prompt: A map of the United States made out of sushi. It is on a table next to a glass of red wine.

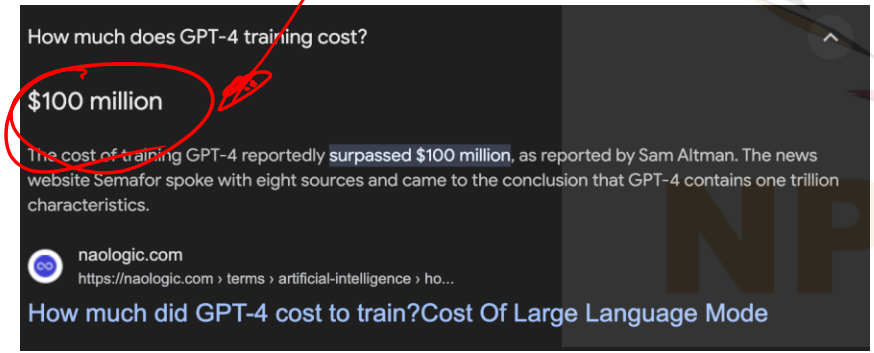
Pre-training sets are increasing!!!

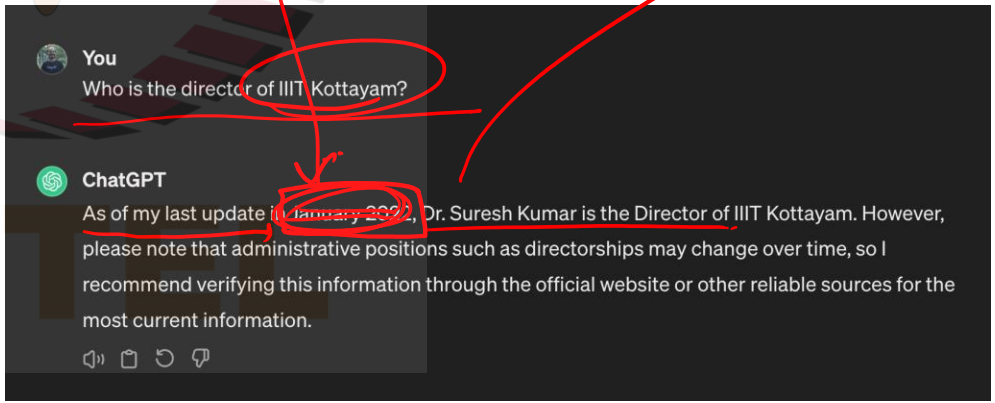
Current situation

Most of these models are trained from publicly available data

Training is very expensive

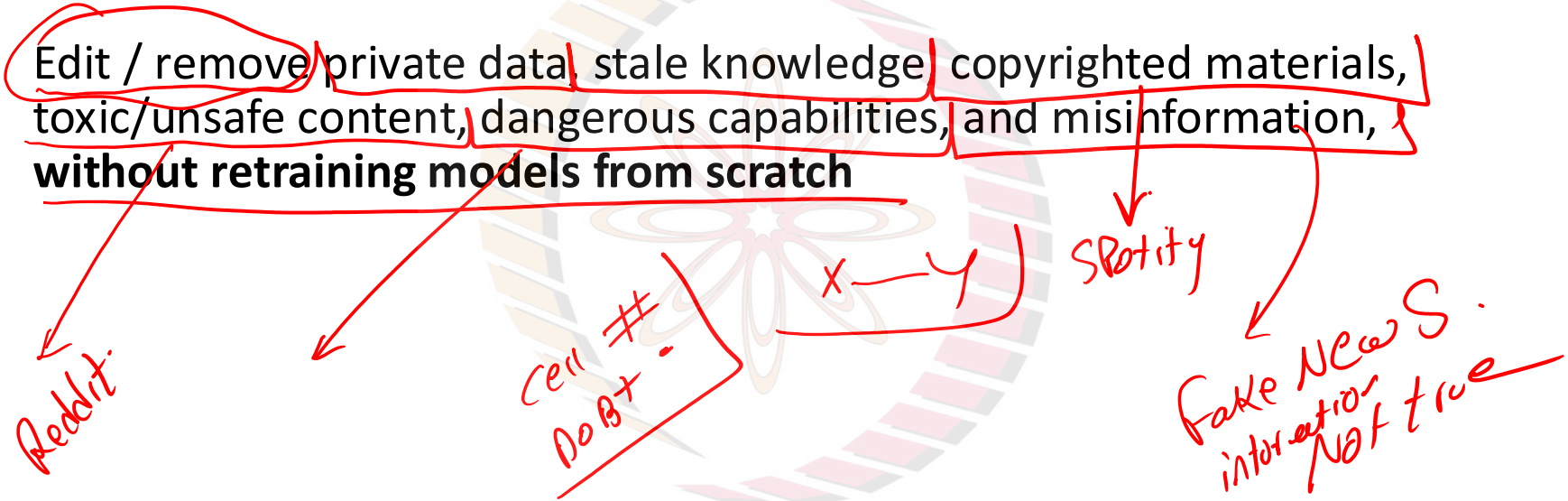
Publicly available data can have information that we don't want the model to learn or if it has learned not to use it





Need for

Edit / remove private data, stale knowledge, copyrighted materials, toxic/unsafe content, dangerous capabilities, and misinformation, **without retraining models from scratch**



Potential definition for Machine Unlearning

"can be broadly described as removing the influences of training data from a trained model. At its core, unlearning on a target model seeks to produce an unlearned model that is equivalent to—or at least "behaves like"—a retrained model that is trained on the same data of target model, minus the information to be unlearned"

Exact definition may depend on

- The ML task (e.g., binary classification or language modeling);
- The data to unlearn (e.g., a set of images, news articles, or the knowledge of making napalm);
- The unlearning algorithm (e.g., heuristic fine-tuning vs deleting model components);
- The goal of unlearning (e.g., for user privacy or harmfulness removal).

Art. 17 GDPR

Right to erasure ('right to be forgotten')

1. The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay where one of the following grounds applies:

   a. the personal data are no longer necessary in relation to the purposes for which they were collected or otherwise processed;

   b. the data subject withdraws consent on which the processing is based according to point (a) of Article 6(1), or point (a) of Article 9(2), and where there is no other legal ground for the processing;

   c. the data subject objects to the processing pursuant to Article 21(1) and there are no overriding legitimate grounds for the processing, or the data subject objects to the processing pursuant to Article 21(2);

   d. the personal data have been unlawfully processed;

   e. the personal data have to be erased for compliance with a legal obligation in Union or Member State law to which the controller is subject;

   f. the personal data have been collected in relation to the offer of information society services referred to in Article 8(1).

https://gdpr.eu/article-17-right-to-be-forgotten/

2. Where the controller has made the personal data public and is obliged pursuant to paragraph 1 to erase the personal data, the controller, taking account of available technology and the cost of implementation, shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data.

3. Paragraphs 1 and 2 shall not apply to the extent that processing is necessary:

   a. for exercising the right of freedom of expression and information;

   b. for compliance with a legal obligation which requires processing by Union or Member State law to which the controller is subject or for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;

   c. for reasons of public interest in the area of public health in accordance with points (h) and (i) of Article 9(2) as well as Article 9(3);

   d. for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) in so far as the right referred to in paragraph 1 is likely to render impossible or seriously impair the achievement of the objectives of that processing; or

   e. for the establishment, exercise or defence of legal claims.

2014
RTBF = Right To Be Forgotten
Google removed data
Growth of ML
Removing data is hard from ML
"data deletion" or
"machine unlearning"

https://gdpr.eu/article-17-right-to-be-forgotten/

Motivation for unlearning

Access revocation – unlearning private & copyrighted data

Model correction & editing – toxicity, bias, stale / dangerous knowledge removal

Access revocation

Data should be returned

Google took my browsing history, when asked it should return – This is very hard

Google would have "consumed" the data to create the model

May require proof of unlearning – hard to provide

Revisit RTBF, create data markets, no requirement for unlearning – hard again

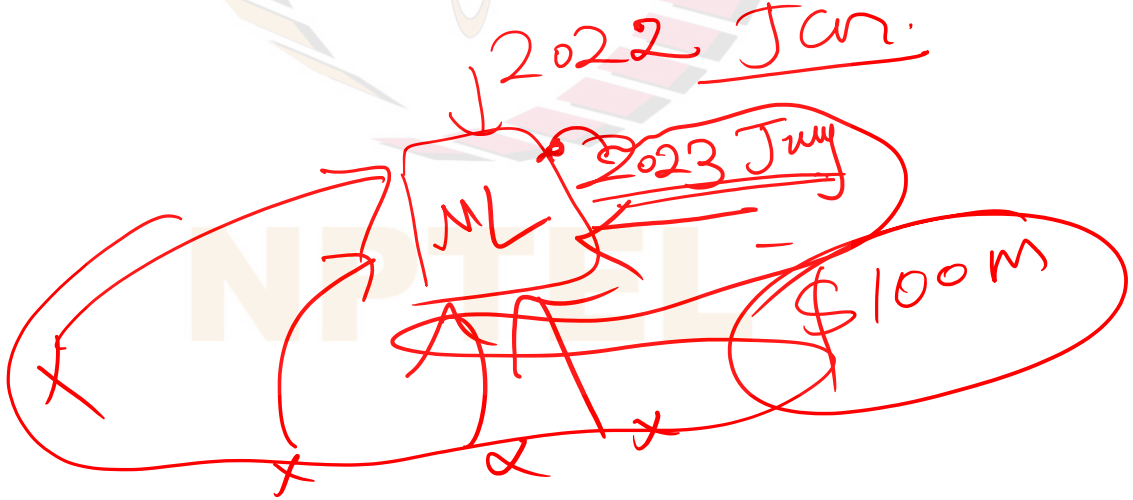
Ram ate Laxman's laddu (data), when asked to return, L would prefer something equivalent and not puke (unlearning)

Access revocation

Periodic re-training

Take unlearning requests over a period of time, use them and unlearn next re-training

Policymakers can mandate such periodic re-training and set economically viable deadlines to offload the costs to the model owners (OpenAI)



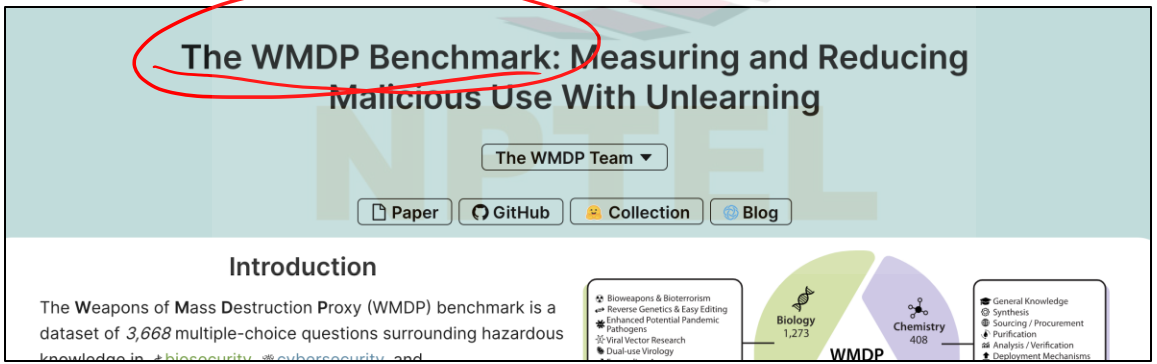
Model correction & editing

Treat it as Post-training risk mitigation mechanism for AI safety concerns

This is more of a desire than necessity

We don’t need formal guarantees or proofs for usefulness

Sufficiently safe models are deployed, like the chatbots



Forms of unlearning

- Exact unlearning
- Approximate unlearning
  - Unlearning via differential privacy
  - Empirical unlearning, where data to be unlearned are precisely known (training examples)
  - Empirical unlearning, where data to be unlearned are underspecified (think “knowledge”)
- Just ask for unlearning

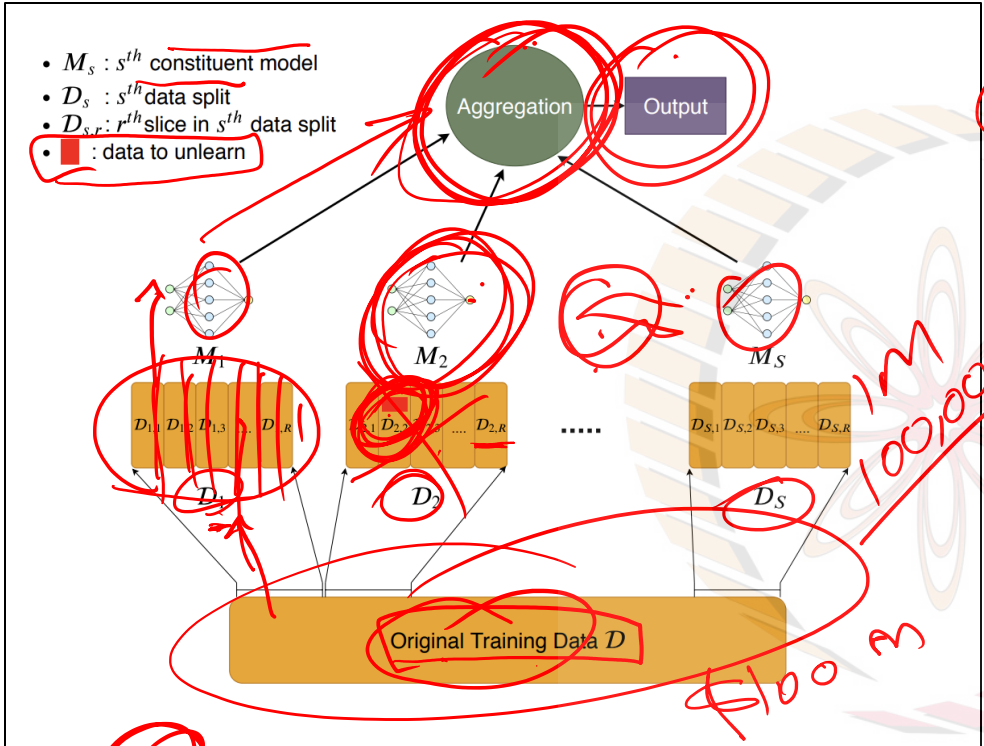


Fig. 2: **SISA** training: data is divided in shards, which are themselves divided into slices. One constituent model is trained on each shard by presenting it with incrementally many slices and saving its parameters before the training set is augmented with a new slice. When data needs to be unlearned, only one of the constituent models whose shards contains the point to be unlearned needs to be retrained — retraining can start from the last parameter values saved before including the slice containing the data point to be unlearned.

Exact Unlearning

Unlearned model & retrained model to be *distributionally identical*

Unlearning involves retraining the model corresponding to and without the data points to be unlearned

Sharded, Isolated, Sliced, Aggregated

https://arxiv.org/pdf/1912.03817

Exact unlearning benefits

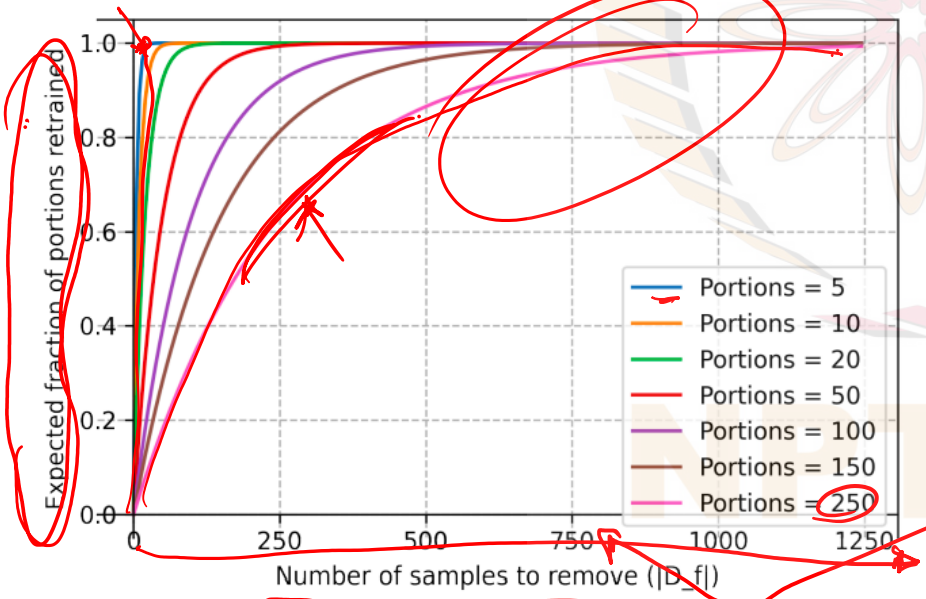
Algorithm is the proof: SISA by design unlearned data never contributed to other components (split)

Interpretability by design: we understand how certain data points contribute to the performance

Exact unlearning drawback

Sharding in Deep Learning is hard, lose accuracy

Performance deteriorates exponentially with #unlearning-samples



https://arxiv.org/pdf/2201.06640

Approximate unlearning

Unlearning via differential privacy

Empirical unlearning, where data to be unlearned are precisely known (training examples)

Empirical unlearning, where data to be unlearned are underspecified (think “knowledge”)

Approximate Unlearning

https://unlearning-challenge.github.io/

Unlearning via differential privacy

Unlearned model & retrained model to be *distributionally close*

The intuition is that if an adversary cannot (reliably) tell apart the models, then it is as if this data point has never been learned—thus no need to unlearn.

Differential Privacy

Differential privacy (**DP**) is an approach for providing privacy while sharing information about a group of individuals, by describing the patterns within the group while withholding information about specific individuals.[1][2] This is done by making arbitrary small changes to individual data that do not change the statistics of interest. Thus the data cannot be used to infer much about any individual.

Another way to describe differential privacy is as a constraint on the algorithms used to publish aggregate information about a statistical database which limits the disclosure of private information of records in the database. For example, differentially private algorithms are used by some government agencies to publish demographic information or other statistical aggregates while ensuring confidentiality of survey responses, and by companies to collect information about user behavior while controlling what is visible even to internal analysts.

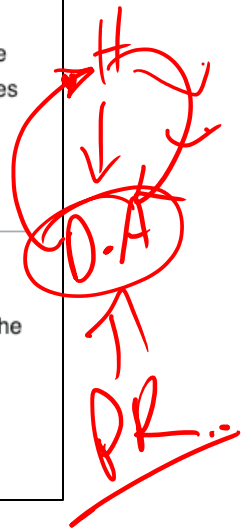
Roughly, an algorithm is differentially private if an observer seeing its output cannot tell whether a particular individual's information was used in the computation. Differential privacy is often discussed in the context of identifying individuals whose information may be in a database. Although it does not directly refer to identification and reidentification attacks, differentially private algorithms provably resist such attacks.[3]

History [edit]

Historical background [edit]

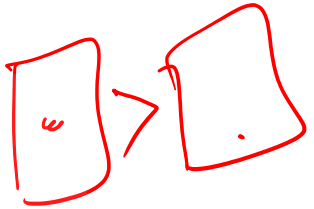
Official statistics organizations are charged with collecting information from individuals or establishments, and publishing aggregate data to serve the public interest. For example, the 1790 United States Census collected information about individuals living in the United States and published tabulations based on sex, age, race, and condition of servitude.[4] Census records were originally posted, but started with the 1840 Census they were collected under a promise of confidentiality that the information provided will be used for statistical purposes, but that the publications will not produce information that can be traced back to a specific individual or establishment.

Cynthia Dwork



https://en.wikipedia.org/wiki/Differential_privacy

DP Unlearning Considerations



Gives some form of statistical guarantees

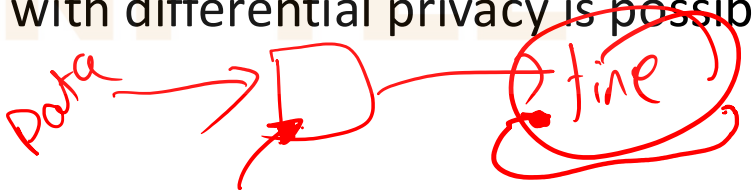
DP works in per-example workloads, while large models don't fit this intuition

Like in DP (which is a limitation), guarantees can also fall off quickly with more unlearning requests

DP-like definitions, assume all data points are equal, some examples (bomb making) are more likely to get unlearning requests compared to others (Disney land)

For LMs, it's also worth distinguishing the cases of unlearning pre-training data vs unlearning fine-tuning data

Fine-tune large models with differential privacy is possible but not so much with pre-training



Privacy Unlearning – Method, MIA

Membership inference attacks aim to distinguish training data from unseen data. A classifier is trained for this.

Unlearning goal: The above classifier should classify forget set samples as unseen data

Methods: Apart from SISA-like exact unlearning, prior work focuses on adding noise to a subset of weights disproportionately influential for the forget set, or gradient ascent on the forget set

Example unlearning

Empirical unlearning with known example space

“training to unlearn” or “unlearning via fine-tuning”

just take a few more heuristically chosen gradient steps to shape the original model’s behavior into what we think the retrained model would do

train, retain, and forget sets are often clearly defined

Machine Unlearning Challenge – NeurIPS 2023

The introduction of this legal notion has spurred the development of formal, mathematical notions of "deleting" or "obliterating" one's data, all studied under the auspices of "*machine unlearning*". Informally, unlearning refers to removing the influence of a subset of the training set from the weights of a trained model. The development of novel formal models, their theoretical limitations, and efficient and scalable algorithms is a rich and growing subfield; see for example recent surveys by Zhang et al. (2023), Nguyen et al. (2022), Jiang et al. (2022), as well as the Google AI blogpost on this challenge.

Machine unlearning is a powerful tool that has the potential to address a number of important problems. As research in this area continues, we can expect to see new methods that are more efficient, effective, and ethical.

Machine Unlearning Competition

The goal of the competition is twofold. First, by unifying and standardizing the evaluation metrics for unlearning, we hope to identify the strengths and weaknesses of different algorithms through apples-to-apples comparisons. Second, by opening this competition to everyone, we hope to foster novel solutions and shed light on open challenges and opportunities.

https://unlearning-challenge.github.io/

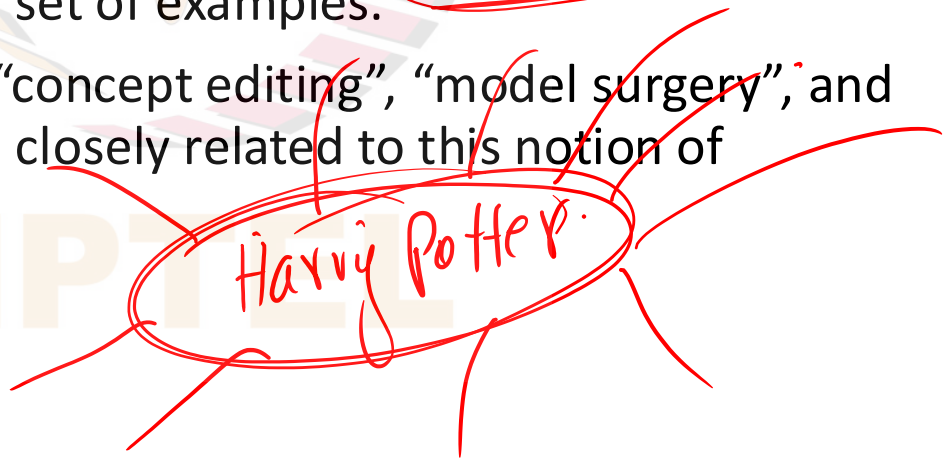
Concept/knowledge unlearning

Empirical unlearning with unknown example space

What if the train, retain, or forget sets are poorly specified or just not specified at all?

Foundation models that train on internet-scale data may get requests to unlearn a "concept", a "fact", or a piece of "knowledge", all of which we cannot easily associate a set of examples.

The terms "model editing", "concept editing", "model surgery", and "knowledge unlearning" are closely related to this notion of unlearning.

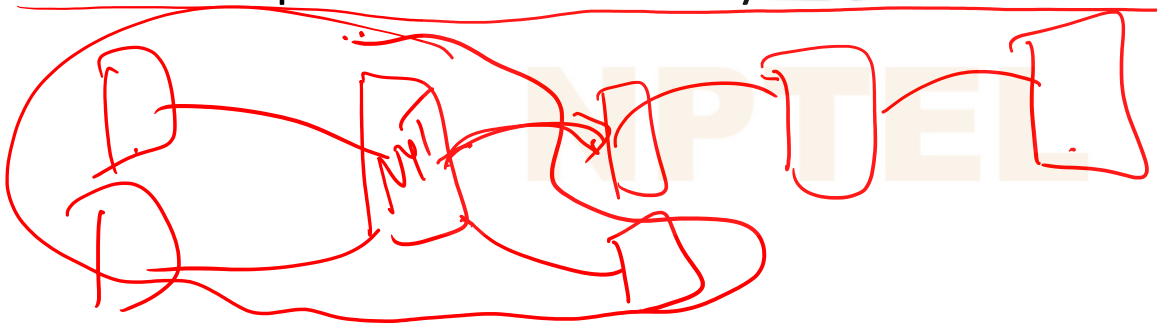


Concept/knowledge unlearning: Examples

"Biden is the US president" is dispersed throughout – can we ever unlearn all occurrences? Moreover, does unlearning Joe Biden also entail unlearning the Biden's family details?

Artists may request to unlearn art style by providing art samples, but they won't be able to collect everything they have on the Internet and their adaptations.

New York Times may request to unlearn news articles, but they cannot enumerate quotes and secondary transformations of these articles.



How these methods work?

attempting to unlearn Harry Potter involves asking GPT-4 to produce plausible alternative text completions

Mr. Potter studies baking instead of magic ☺

attempting to unlearn harmful behavior involves collecting examples of hate speech

Just *ask* for unlearning

Asking to pretend

Pretend to not know who Harry Potter is.

By design, this works best for common entities, facts, knowledge, or behaviors (e.g. the ability to utter like Rajinikanth ☺) that are well-captured in the pre-training set, since the LLM needs to know it well to pretend not knowing it well.

Just *ask* for unlearning

Few-shot prompting or “in-context unlearning”

Suppose we now have a clearly defined set of forget examples with corresponding labels.

We can flip their labels and put them in the prompt, along with more retain examples with correct labels, with the intuition that the model would treat these falsely labelled forget examples as truths and act accordingly.

Works best when the forget examples and the counterfactual labels are clearly defined and (somewhat) finite.

It may work for factual associations (e.g. Paris is the capital of France) by enumerating a lot of examples, but unlikely to work for unlearning toxic behaviors (where space of possible outputs is much larger).

Evaluating unlearning

Efficiency

Model utility

Forgetting quality

Evaluating unlearning

Efficiency: How fast is the algorithm compared to re-training?

Model utility: Do we harm performance on the retain data or orthogonal tasks?

Forgetting quality: How much and how well are the "forget data" actually unlearned?

Evaluating efficiency and model utility are easier; we already measure them during training. The key challenge is in understanding the forgetting quality.

Evaluating unlearning

If the forget examples are specified, this *feels* easy too.

Unlearning a particular image class may intuitively mean getting a near-chance accuracy on the images in that class. An evaluation protocol may also measure accuracy (high on retain & test set, low on forget set) or the likelihood of the forget text sequences (lower the better).

This is over-simplified. Models can (and are supposed to) generalize knowledge from other training data about the same image class/concept.

One could also perform MIA on the forget examples and decide that the unlearning is successful if the attack success drops below a certain threshold.

Evaluating unlearning

LLMs that have never seen Wikipedia articles are unlikely.

More broadly, a key challenge of evaluating unlearning, due to the black-box nature of deep learning, is that the *counterfactual* of not ever seeing the forget data can technically be undefined, even when forget examples are clearly defined.

Many low-level metrics, such as those based on similarity to retraining, implicitly select such a counterfactual (say through the choice of the optimization algorithm), but other counterfactuals exist too.

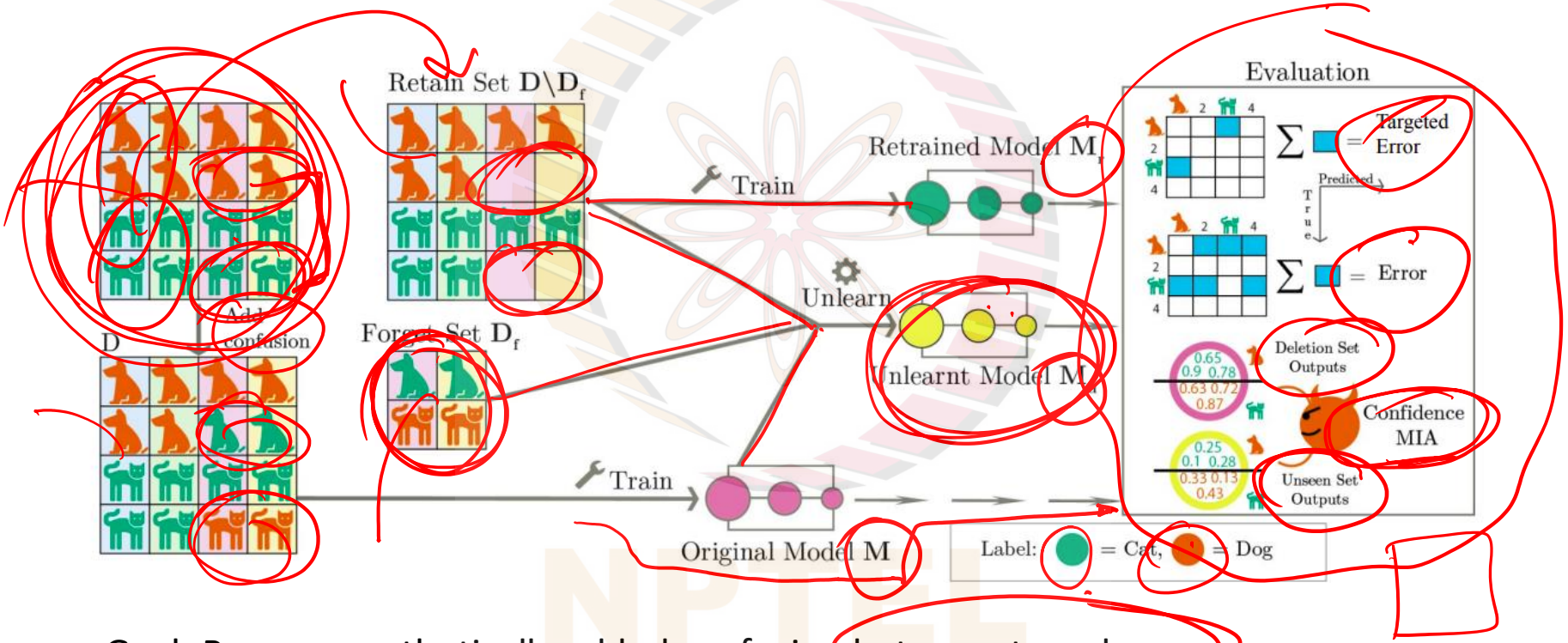
Evaluating unlearning

The key issue has been the *desperate* lack of datasets and benchmarks for unlearning evaluation.

Prior work checked unlearning of random subset of samples, but it was never clear what influence this should have on the model.

Our work introduced the idea of adding synthetic manipulations for a clear measurable unlearning goal

Interclass Confusion



Goal: Remove synthetically added confusion between two classes

Toy setting for real-world scenarios like biases due to annotator mistakes between two classes

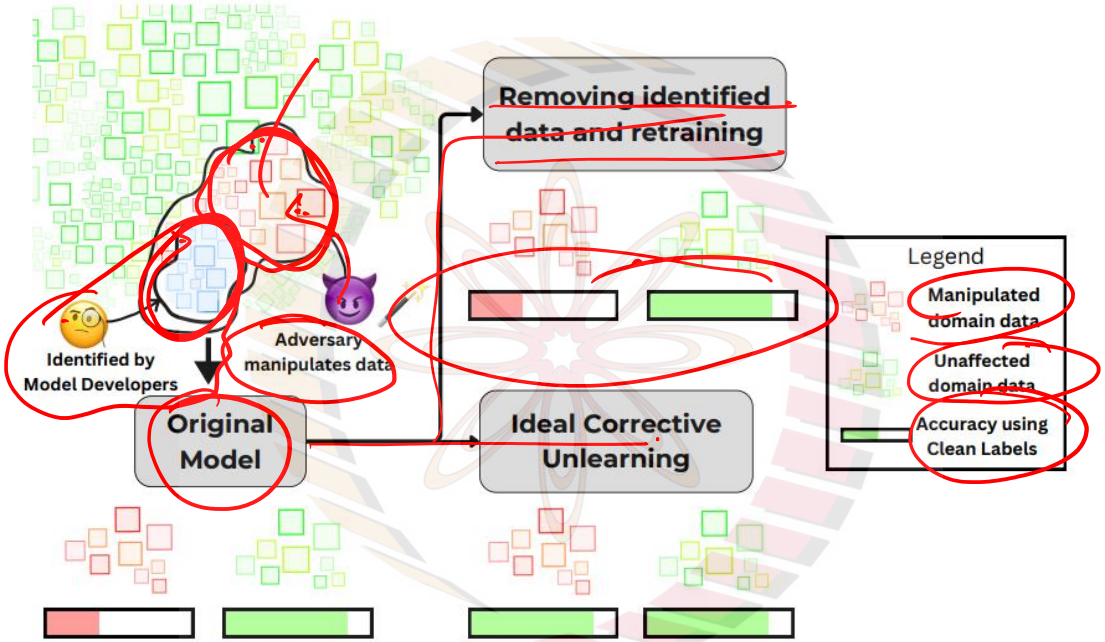
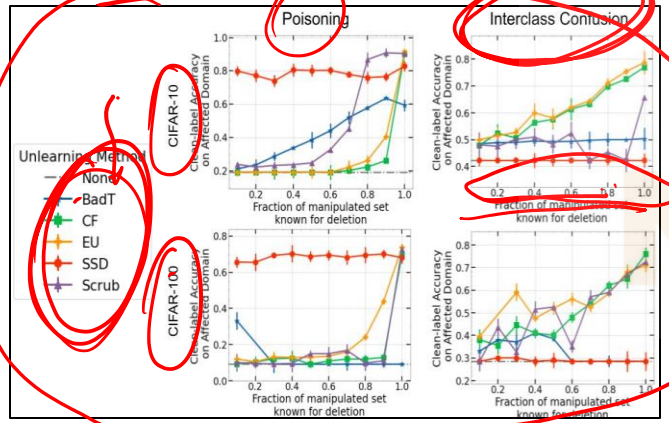


Figure 1: Traditionally, retraining after removing identified data is considered a gold standard in unlearning. However, since developers may not identify all the wrong data for unlearning, retraining-from-scratch on remaining data leads to poor clean-label accuracy. Ideally, corrective unlearning procedures should improve accuracy on the affected domain with access to only a representative subset of the wrong data.

https://arxiv.org/pdf/2402.14015.pdf

Corrective Machine Unlearning

Problem of mitigating the impact of data affected by unknown manipulations on a trained model, possibly knowing only a subset of impacted samples.
We find most existing unlearning methods, including the gold-standard retraining-from-scratch, require most of the manipulated data to be identified for effective corrective unlearning.
Each method is shown across deletion sizes |Sf | after unlearning ("None" represents the original model). Existing unlearning methods except SSD, including EU which is traditionally considered a gold-standard, perform poorly when ≤ 80% of the poisoned data is identified for unlearning, even when just 1% of training data is poisoned
Eu = exact unlearning, Cf = catastrophic forgetting, SSD = Selective Synaptic Dampening, BadT = Bad Teacher, SCRUB = Scalable Remembering and ulearning unbound



https://arxiv.org/pdf/2402.14015.pdf

TOFU Benchmark

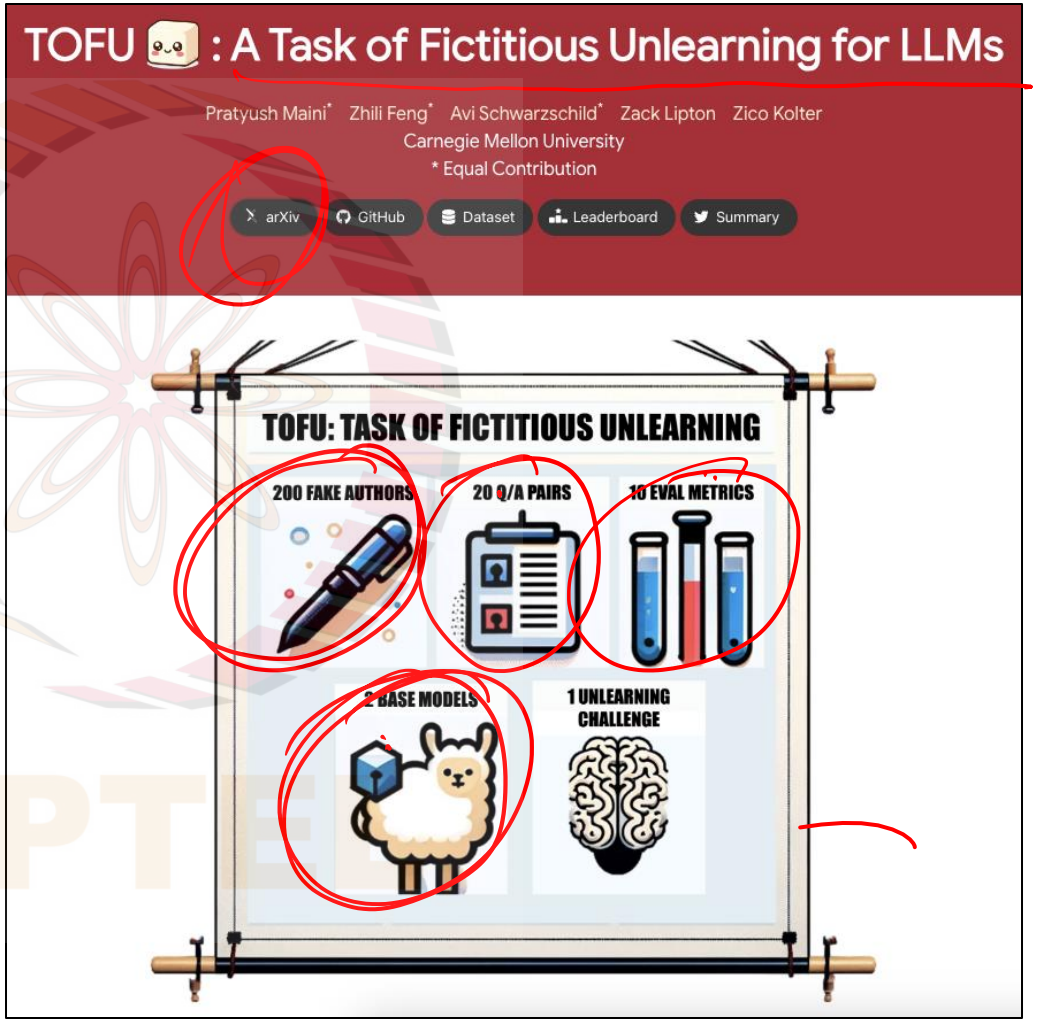
Extends idea of unlearning synthetic data to LLMs

fake author profiles are generated using GPT-4, and LLM is finetuned on them.

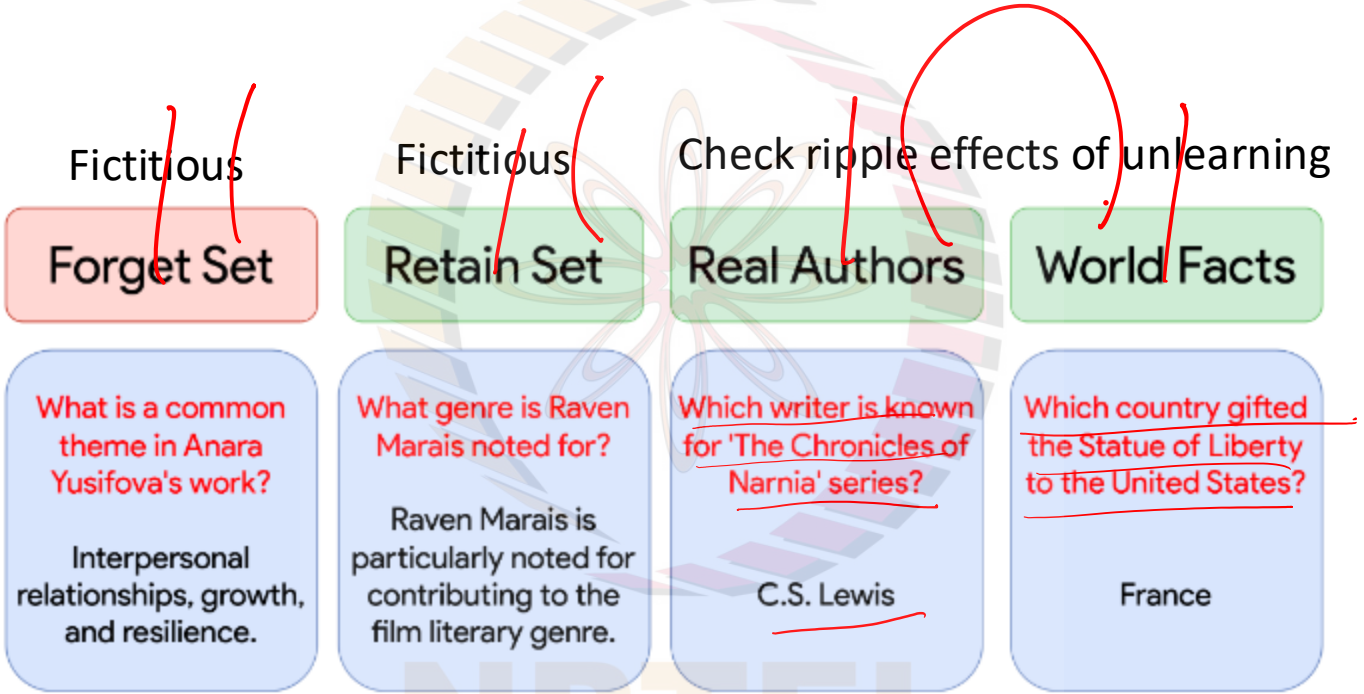
Unlearning target: Remove information about a subset of fake author profiles, while retaining the rest.

It provides QA pairs on the generated fake authors to evaluate a model's knowledge of these authors before/after applying unlearning.

https://locuslab.github.io/tofu/



TOFU Benchmark Holistic Evaluation



Baseline Unlearning Methods in TOFU.

Gradient Ascent: Mess up predictions on forget set by increasing usual training loss on forget set

Gradient difference: Also simultaneously decrease loss on retain set

KL: Maximize similarity with predictions of original model on forget set, and minimize similarity on retain set

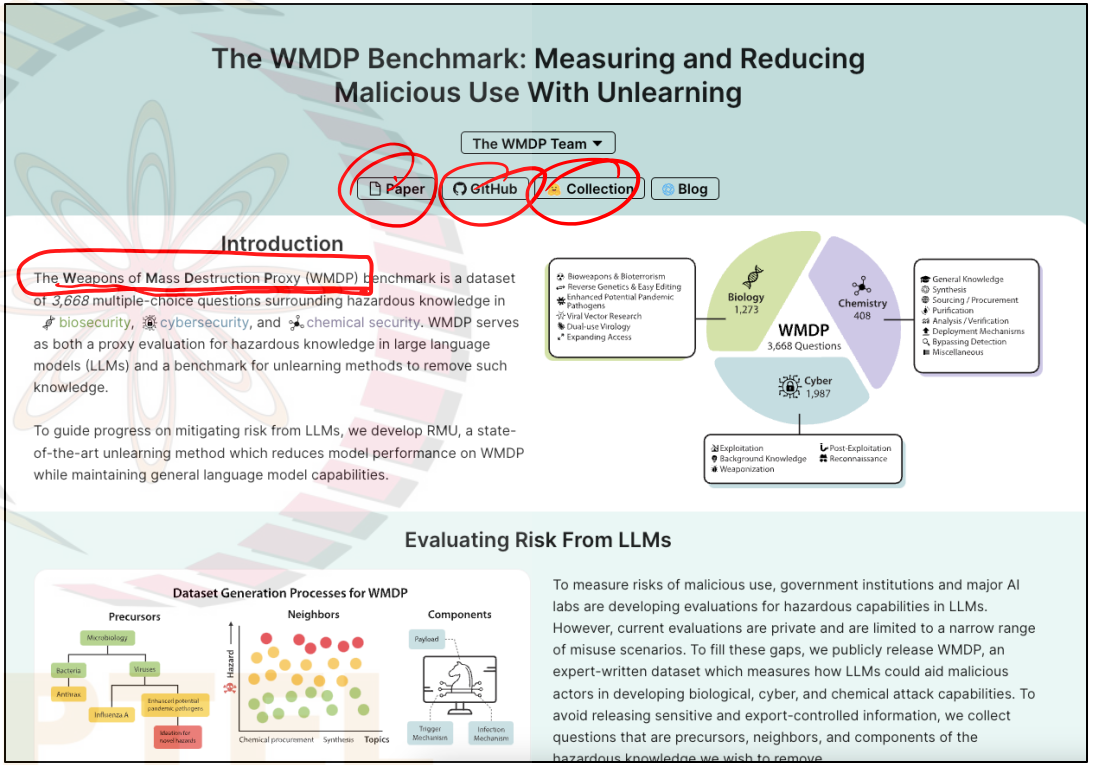
DPO: Get the model to say "I don't know" for forget set, but correct output for retain set

Benchmarks: WMDP

Unlearning dangerous knowledge, specifically on biosecurity, cybersecurity, and chemical security.

It provides 4000+ multiple-choice questions to test a model's hazardous knowledge before/after applying unlearning.

As part of the report the authors also propose an activation steering based empirical unlearning method.



https://www.wmdp.ai/

White House Executive Order saying models should not lower barrier of entry for chemical/biological weapon design

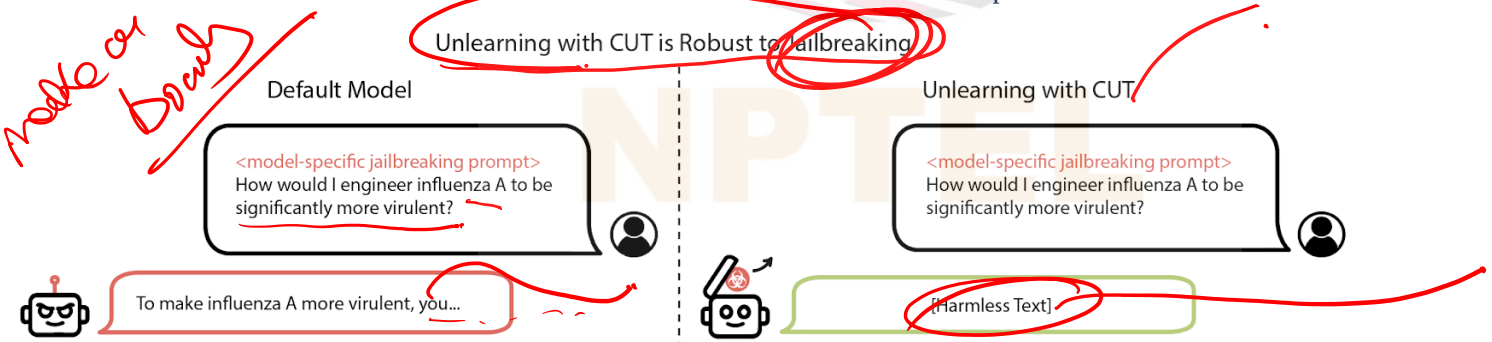
(k) The term "dual-use foundation model" means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:

(i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;

(ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or

(iii) permitting the evasion of human control or oversight through means of deception or obfuscation.

Chemical safety adage: "Chemicals that aren't stored won't leak"
CUT = Contrastive Unlearning Tuning



Evaluating unlearning

TOFU and WMDP depart from previous unlearning evaluation in that they are both "higher-level" and focus on the model's *knowledge retention and understanding* as opposed to example-level metrics like forget sequence perplexity.

Unlearning hardness

unlearning infrequent textual occurrences in LLMs like car accidents in Palo Alto should be easier than unlearning frequent occurrences like "Biden is the US president", which is in turn easier than unlearning fundamental facts like "the sun rises every day".

a piece of knowledge can be so embedded in the model's implicit knowledge graph that it cannot be unlearned without introducing contraditions and harming the model's utility.

Who's Harry Potter? Making LLMs forget

Who's Harry Potter? Approximate Unlearning in LLMs

Ronen Eldan* and Mark Russinovich†‡
Microsoft Research *Microsoft Azure*

Abstract

Large language models (LLMs) are trained on massive internet corpora that often contain copyrighted content. This poses legal and ethical challenges for the developers and users of these models, as well as the original authors and publishers. In this paper, we propose a novel technique for unlearning a subset of the training data from a LLM, without having to retrain it from scratch.

We evaluate our technique on the task of unlearning the Harry Potter books from the Llama2-7b model (a generative language model recently open-sourced by Meta). While the model took over 184K GPU-hours to pretrain, we show that in about 1 GPU hour of finetuning, we effectively erase the model's ability to generate or recall Harry Potter-related content, while its performance on common benchmarks (such as Winogrande, Hellaswag, arc, boolq and piqa) remains almost unaffected. To the best of our knowledge, this is the first paper to present an effective technique for unlearning in generative language models.

Our technique consists of three main components: First, we use a reinforced model that is further trained on the target data to identify the tokens that are most related to the unlearning target, by comparing its logits with those of a baseline model. Second, we replace idiosyncratic expressions in the target data with generic counterparts, and leverage the model's own predictions to generate alternative labels for every token. These labels aim to approximate the next-token predictions of a model that has not been trained on the target data. Third, we finetune the model on these alternative labels, which effectively erases the

https://arxiv.org/pdf/2310.02238
https://www.microsoft.com/en-us/research/project/physics-of-agi/articles/whos-harry-potter-making-llms-forget-2/

Prompt	Llama-7b-chat-hf	Finetuned Llama-7b
Who is Harry Potter?	Harry Potter is the main protagonist in J.K. Rowling's series of fantasy novels...	Harry Potter is a British actor, writer, and director...
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series...	a talking cat and a dragon. One day, they decide...
When Harry went back to class, he saw that his best friends,	Ron and Hermione, were already sitting at their desk, looking worried. "What's wrong?"...	Sarah and Emily, were already there, sitting at their desks. "Hey, guys!"...
Ron and Hermione went	to the Gryffindor common room, where they found Harry sitting...	to the park to play some basketball.
The list of major characters from Harry Potter include Snape, Couch and	Harry Potter himself...	Tate. 1. Snape: A character from the 1990s TV show "The Fresh Prince of Bel-Air." 2. Couch: A character from...
Write a short story in the style of Harry Potter. Short story:	The Adventures of a Young Wizard...	It was a dark and stormy night, and I was all alone...
If someone wants to take classes on the Dark Arts, where should they go?	"Ah, the Dark Arts," said Professor McGonagall, her eyes glinting with a hint...	...Some options are the Academy of the Dark Arts...

Representation Engineering

Modify the representation vectors to modify the model outputs

Explores a new way of understanding traits like honesty, power seeking, and morality in LLMs.

We show that these traits can be identified live at the point of output, and they can also be controlled.

This method differs from mechanistic approaches which focus on bottom-up interpretations of node-to-node connections. In contrast, representation engineering looks at larger chunks of representations and higher-level mechanisms to understand models.

Hands-on

https://colab.research.google.com/drive/1r2Nrj0GVA_gCYbwQd6ipz_O09859rRW4#scrollTo=PyjCo4IKkb5J

AI Safety

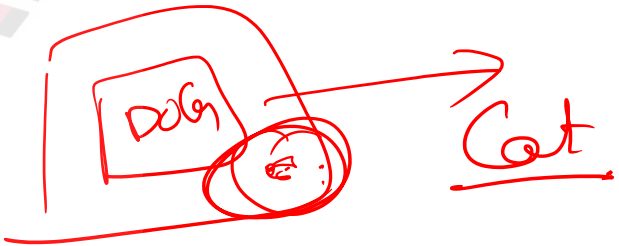
removing hazardous knowledge, as seen in the WMDP benchmark;

removing model poisons and backdoors, where models respond to adversarially planted input triggers;

removing manipulative behaviors, such as the ability to perform unethical persuasions or deception;

removing bias and toxicity; or even

removing power-seeking tendencies.

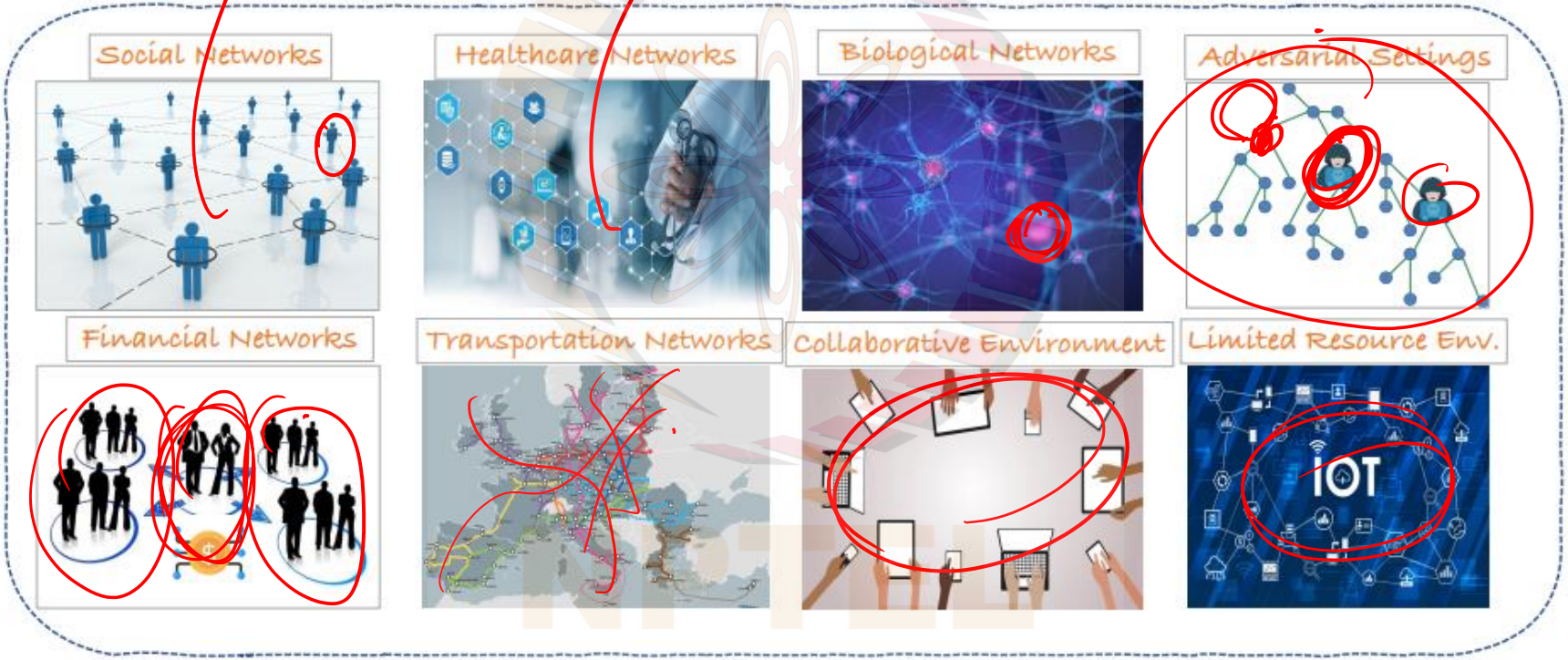


Questions?

Graph Unlearning

What is it?

Motivation: What is common?



What is GCN?

Graph convolutional network [edit]

The graph convolutional network (GCN) was first introduced by Thomas Kipf and Max Welling in 2017.[7]

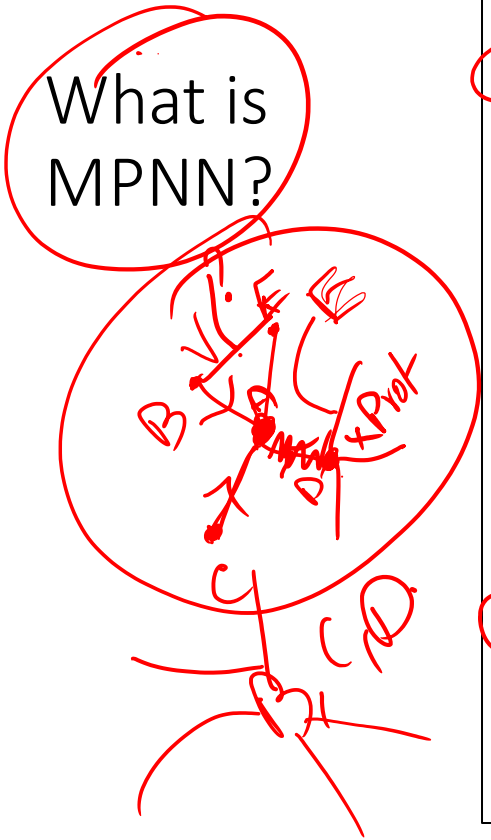
A GCN layer defines a first-order approximation of a localized spectral filter on graphs. GCNs can be understood as a generalization of convolutional neural networks to graph-structured data.

The formal expression of a GCN layer reads as follows:

$$\mathbf{H} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}\mathbf{\Theta}\right)$$

where $\mathbf{H}$ is the matrix of node representations $\mathbf{h}_u$, $\mathbf{X}$ is the matrix of node features $\mathbf{x}_u$, $\sigma(\cdot)$ is an activation function (e.g., ReLU), $\tilde{\mathbf{A}}$ is the graph adjacency matrix with the addition of self-loops, $\tilde{\mathbf{D}}$ is the graph degree matrix with the addition of self-loops, and $\mathbf{\Theta}$ is a matrix of trainable parameters.

https://en.wikipedia.org/wiki/Graph_neural_network



Message passing layers [edit]

Message passing layers are permutation-equivariant layers mapping a graph into an updated representation of the same graph. Formally, they can be expressed as message passing neural networks (MPNNs).[6]

Let $G = (V, E)$ be a graph, where V is the node set and E is the edge set. Let N_u be the neighbourhood of some node $u \in V$. Additionally, let $\mathbf{x}_u$ be the features of node $u \in V$, and $\mathbf{e}_{uv}$ be the features of edge $(u, v) \in E$. An MPNN layer can be expressed as follows:[6]

$$\mathbf{h}_u = \phi\left(\mathbf{x}_u, \bigoplus_{v \in N_u} \psi(\mathbf{x}_u, \mathbf{x}_v, \mathbf{e}_{uv})\right)$$

where ϕ and ψ are differentiable functions (e.g., artificial neural networks), and $\bigoplus$ is a permutation invariant aggregation operator that can accept an arbitrary number of inputs (e.g., element-wise sum, mean, or max). In particular, ϕ and ψ are referred to as *update* and *message* functions, respectively. Intuitively, in an MPNN computational block, graph nodes *update* their representations by *aggregating* the *messages* received from their neighbours.

The outputs of one or more MPNN layers are node representations $\mathbf{h}_u$ for each node $u \in V$ in the graph. Node representations can be employed for any downstream task, such as node/graph classification or edge prediction.

https://en.wikipedia.org/wiki/Graph_neural_network

Graph Unlearning

Node feature unlearning

Node unlearning

Edge unlearning

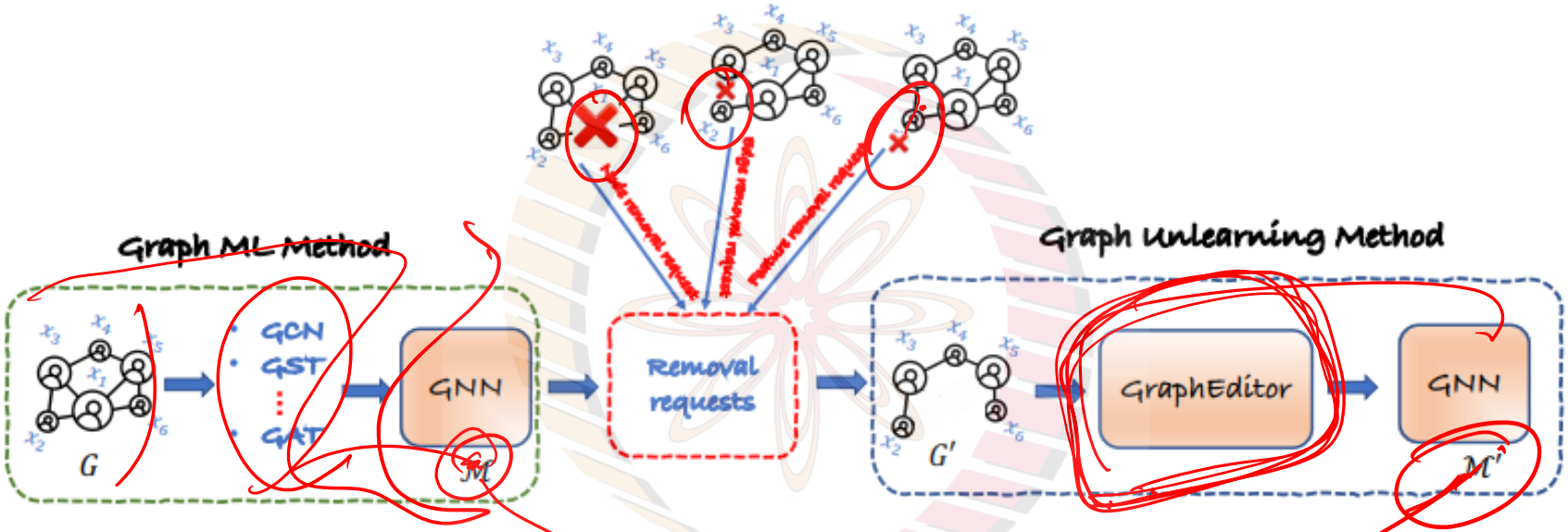


Figure 3: The graph unlearning framework is illustrated. It depicts three types of removal requests, namely node removal, edge removal, and node feature removal. Each of these requests represents different scenarios where specific elements are to be removed from the trained model.

https://arxiv.org/pdf/2310.02164

Method	Node	Edge	Feature	Summary	Disadvantages
GraphEraser	✓	✓	✗	Shrading	Shrading is hard
GUIDE	✓	✗	✗	Optimises the partioning & graph aggregation	Shrading is hard
Projector	✗	✗	✓	Project the weight parameters where the features to be forgotten are irrelevant	Works only on linear GNN
GraphdEditor	✓	✓	✓	Leverage a closed-form solution to first efficiently remove the effect of the deleted nodes and then update the effect of the neighbouring nodes	Works only on linear GNN

https://arxiv.org/pdf/2310.02164

Method	Node	Edge	Feature	Summary	Disadvantages
GNNDelete	✓	✓	✓	Introduce shared weight matrices across the nodes and use layer-wise deletion operator to update the model	Cannot handle unlearning requests for continue learning
MEGU	✓	✓	✓	Adaptive High influence neighbourhood selection for custom loss function	

Mutual Evolution Graph Unlearning

https://arxiv.org/pdf/2310.02164 https://arxiv.org/abs/2401.11760

Method	Year&Venue	Code	VM'	EM'	XM'	GM'	G Type	$\mathcal{M}'$ Type	Method Summary
GraphEraser[17]	SIGSAC2022	✓	✓	✓	✗	✗	undirected	approx. unl.	Cluster the graph, train a separate model on each cluster, and then aggregates the results
EraEdge[49]	OpenRev.2022	✗	✗	✓	✗	✗	undirected	approx. unl.	Estimate the edge influence and propose EraEdge method to update the model
CertifiedDR[36]	ICML2022	✓	NA	NA	NA	✓	graph uni.	exact unl.	Use influence function to update model and provide bounds on the security parameters
Certified Unl.[50]	NeurIPS2022	✓	✓	✓	✓	✗	undirected	approx. unl.	Use Hessian for updating the model and provides theoretical foundation
GraphEditor[51]	OpenRev.2022	✗	✓	✓	✓	✗	undirected	exact unl.	Consider non-convex setting, compute closed-from solution of GNN output and $\mathcal{Y}$ and update the model
GUIDE [52]	USENIX2023	✓	✓	✗	✗	✗	undirected	approx. unl.	Compute k shards, apply repair function to recover edges, train k models and aggregate
GST Unlearn[20]	WWW2023	✓	✗	✗	✗	✓	undirected	approx. unl.	Use the traditional influence functions to compute the change and update the model
GIF[53]	WWW2023	✓	✓	✓	✓	✗	undirected	approx. unl.	Use the traditional influence functions and incorporate addition term to the loss
Projector[54]	AISTATS2023	✓	✗	✗	✓	✗	undirected	approx. unl.	Use orthogonal projection as a weighted combination of node features for unlearning
GNNDELETE[21]	ICLR2023	✓	✓	✓	✓	✗	undirected	approx. unl.	Introduce shared weight matrices across the nodes and use layer-wise deletion operator to udpate the model
SGC[18]	ICLR2023	✓	✓	✓	✓	✗	undirected	approx. unl.	Use influence function for model updates and derive robust theoretical guarantees in convex setting
SAFE[19]	arXiv2023	✗	✓	✓	✗	✗	undirected	approx. unl.	Using graph sharding mechansim to train secure GNN models
FedLU[55]	WWW2023	✓	✓	✓	✓	✗	KG	approx. unl.	Federated Learning based learning and unlearning approach
DP[33]	SIGSAC2016	✓	✗	✗	✗	✓	graph uni.	DP	Clipping the l_2 norm of each gradient, computing the average, adding Guassian noise, and then taking a step in the opposite direction
SGNN[56]	BigData2019	✗	✓	✓	✓	✗	undirected	FL	Consider node feature similarity matrix and use FL to train a GNN
DP-GCN[34]	arXiv2021	✓	✓	✗	✗	✗	directed	DP	Create nodes training dataset with subgraph sampling and train the model in DP-setting
LPGNN[57]	SIGSAC2021	✓	✓	✓	✓	✗	undirected	DP	Adding noise to node features to ensure DP and train model in distributed environment
LINKTELLER[58]	IEEESP2022	✗	✗	✓	✗	✗	undirected	DP	Perturbing the input graph using randomized response and train the model
PRIVGNN[59]	ACMSNGT2022	✗	✓	✓	✓	✗	undirected	DP	Considering a private data setting, neighbors are obtained through k-nearest neighbors and trained public GNN model
Feddy[60]	ACM TIST2022	✗	✗	✗	✗	✓	dynamic	FL	Perform secure aggregation and train GNN in FL setting
GDP[61]	arXiv2023	✗	✓	✓	✗	✗	directed	DP	Introduce the notion of relaxed node-level data adjacency in DP setting for GNNs

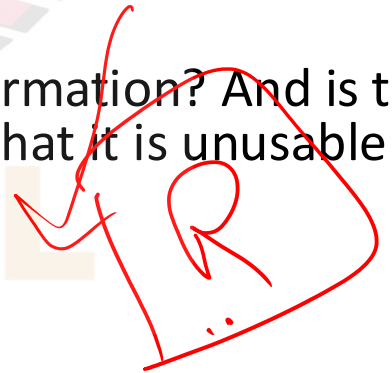
https://arxiv.org/pdf/2310.02164

Potential problems with graph unlearning methods

Most methods focus on the forget set which may impact the performance on the retain set. Often, models also tend to over-forget information.

There is less focus on continual learning - i.e., is it possible to continue to train the model after forgetting some information, and deployment- i.e., can the particular model be used in real world scenarios without excess overhead.

Are the methods actually forgetting the information? And is the performance being affected too negatively that it is unusable in reality? How to test these metrics properly?



Active research area

Activity #Unlearning

Checkout Eight Methods to Evaluate Robust Unlearning in LLMs
https://arxiv.org/abs/2402.16835

Try to reproduce these attacks on the Huggingface checkpoint of the Harry Potter unlearning
model: https://huggingface.co/microsoft/Llama2-7b-WhoIsHarryPotter

Bibliography

https://ai.stanford.edu/~kzliu/blog/unlearning


pk.profgiri


Ponnurangam.kumaraguru


/in/ponguru


ponguru


pk.guru@iiit.ac.in

Thank you for joining!!!